音響学会秋季研究発表会 3-6-3

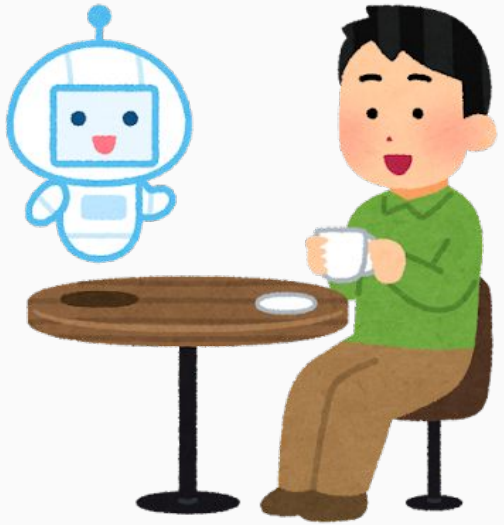
J-CHAT: 音声言語モデルのための大規模日本語対話音声コーパス

◎中田亘，関健太郎，谷中瞳，齋藤佑樹（東大）
高道慎之介（慶大，東大），猿渡洋（東大）

MOONSHOT RESEARCH & DEVELOPMENT PROGRAM
AVATAR SYMBIOTIC SOCIETY

背景：音声対話システム

- 人-AIのコミュニケーションを円滑化
  - 笑い声，叫び声，非言語発話を再現する
- 音声言語モデル（Spoken Language Model)
  - 音声対話システムを実現する上で重要
  - 音声の言語的構造をモデル化
  - 大量の音声データを要する



目標：表現力豊かな音声対話に向けて

現状の主な課題（以下の２点）の解決

1. 大規模かつ多様なデータセットの構築
    a. タスクがTTSなどと比べ難しい，大きなデータセットが必要
    b. 既存の対話データセットは最大でも2000時間
2. 対話を適切にモデリングする手法の検討
    a. そもそも十分なデータセットがない → 研究が進んでいない
    b. 先行研究[Nguyen+22]でもデータセットの小ささを指摘

関連研究

音声対話データセット: Fisher corpus[Cieri+04]:

- 自発的な電話音声を収録
- 2000時間の英語対話音声　対話音声資源として最大
- 有償（USD 7,000)

対話音声言語モデル: dGSLM[Nguyen+22]

- 対話音声のみから言語モデルを学習
- ターンテイキングや笑い声などの生成を確認
- データセットが小さいため，有意味性が低い

関連研究：一方その頃BigTechでは....

Google, Soundstorm[Borsos+23]

- 10万時間の対話音声😓

OpenAI, Whisper [Radford+23]

- 68万時間の音声😱

Meta, Seamless [Barrault+23]

- 450万時間の音声🤯

大規模内製データセットでモデル開発
データセットは非公開＆詳細は謎

BigTech以外は
研究の土俵にすら
上がれない！

J-CHAT：約７万時間の日本語対話音声コーパス

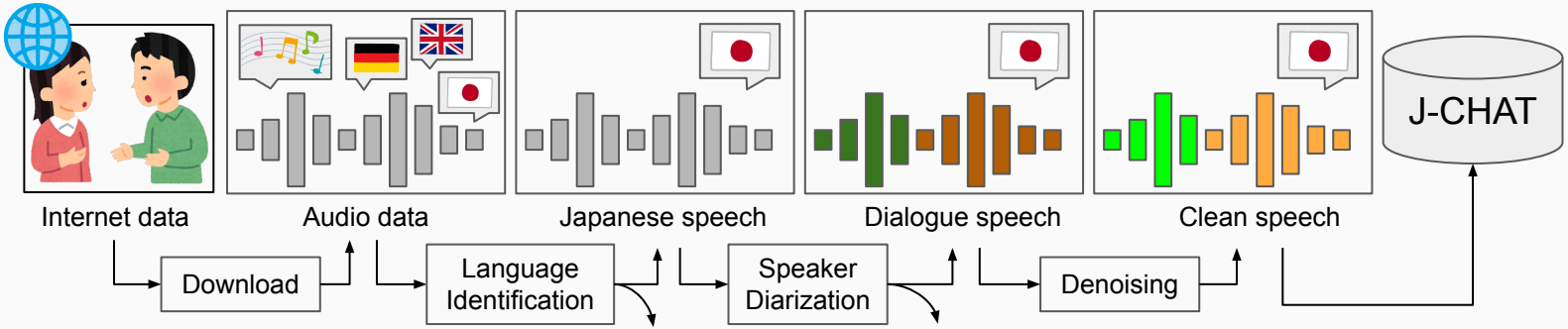
💥対話音声データセットとして最大のコーパス

😎日本語で最も大きな音声コーパス

🤑データは公開済み，非商用なら誰でも使用可能

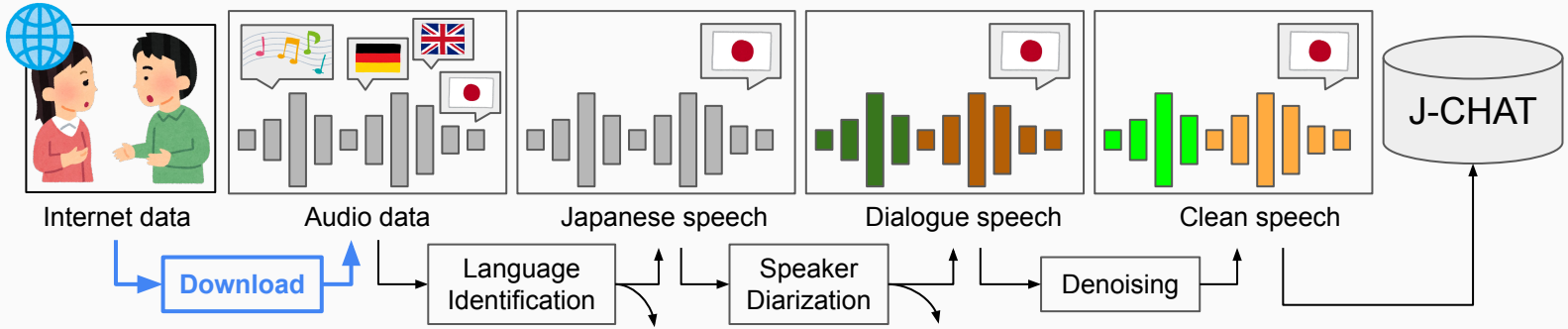
Corpus name	Size(hours)	Open-source	Dialogue	Spontaneous	Clean speech
STUDIES (Saito et al., 2022)	8.2	✅	✅		✅
DailyTalk (Lee et al., 2023)	20	✅	✅	✅	✅
Fisher (Cieri et al., 2004)	$2k$		✅	✅	✅
GigaSpeech (Chen et al., 2021)	$33k$	✅		-	
J-CHAT (This study)	$69k$	✅	✅	✅	✅

データセット収集手法（概要）



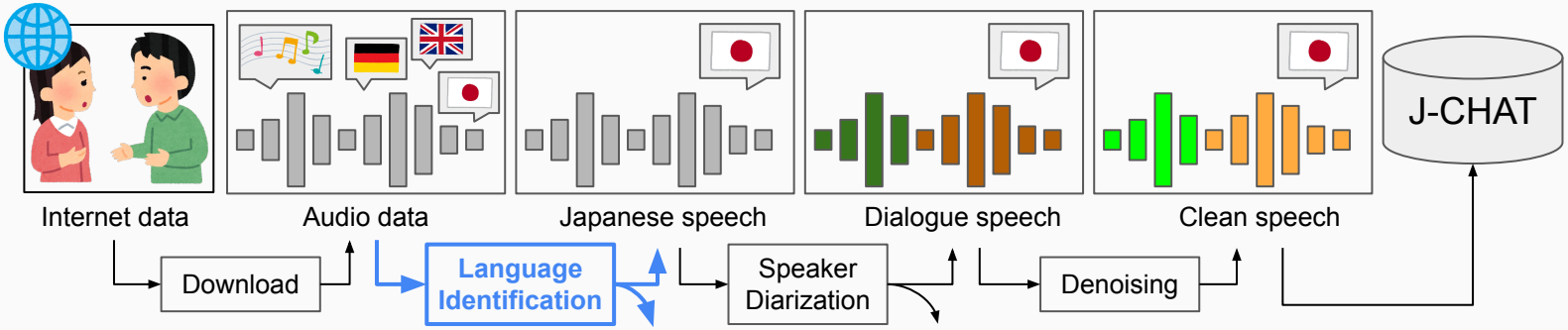
※詳細は以降のスライドで説明
データはYouTube及びPodcastから収集
言語識別，話者ダイアライゼーションにより日本語音声対話を抽出

Step1. Download



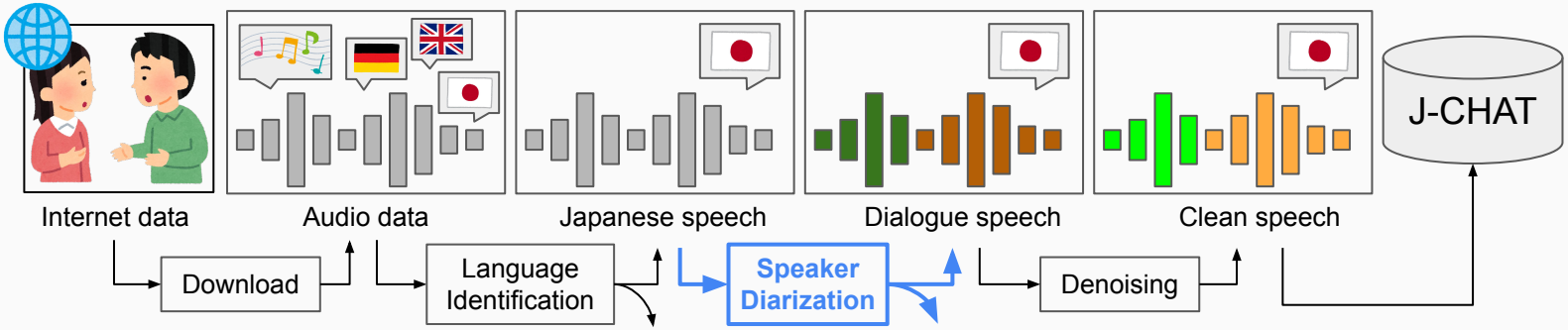
YouTube: ランダムな単語でYouTube検索-> Download
Podcast: ポッドキャストチャンネルのRSSを収集 -> Download

Step2. Language Identification



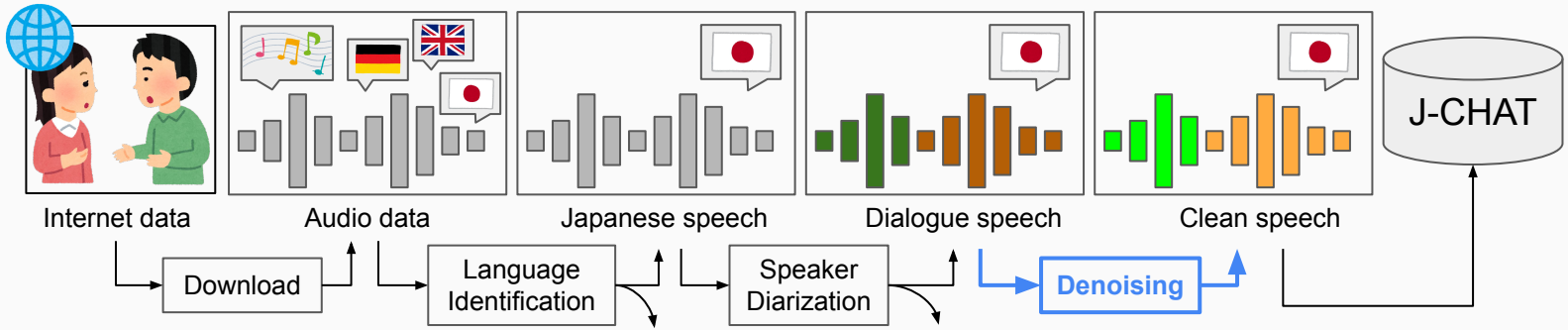
Whisper[Radford+23]を用いて日本語を抽出
YouTube: 55.7% Podcast 84.7% を抽出

Step3. Speaker Diarization



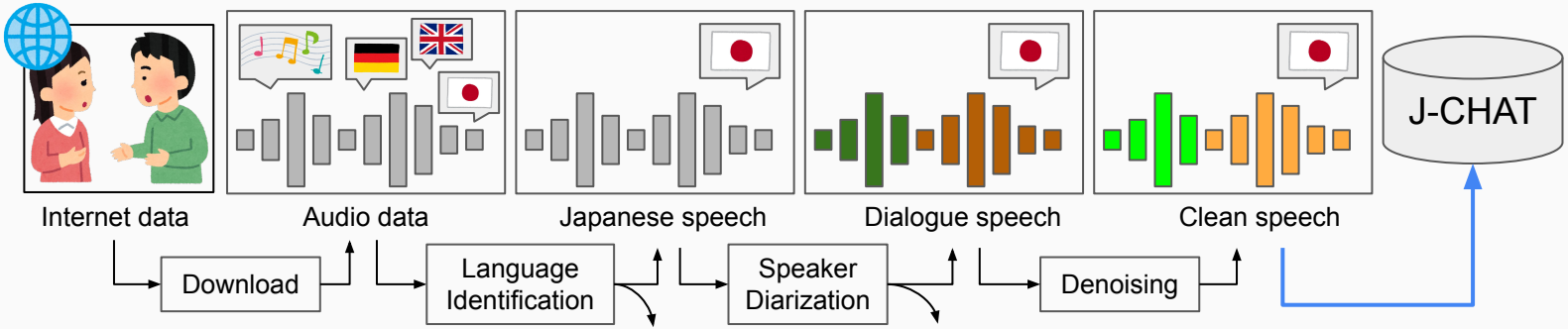
「誰が，いつ喋っているか」を推定するDiarizationを実施
発話区間の80%以上を一人の話者が占めているデータ（独話）を除去
YouTube 41.9% Podcast 45%を抽出
5秒以上発話がない場合は対話が切れたとして扱う

Step4. Denoising



demucs[Rouard+23] を用いてBGMなどのノイズを除去

Step5. 完成！



以上のプロセスでコーパスが完成

- 音声
- 対話ラベル：誰がいつ喋ってるか（Speaker Diarization で付与）
  - 例）ID 1 さんが 0:02〜0:05; ID 2 さんが0:08〜0:20; ....

最終的に残った音声のサンプル

YouTube

Podcast

J-CHATの大きさ

Podcastの方が対話数及び平均対話持続時間が長い
話者数はYouTubeの方が多い

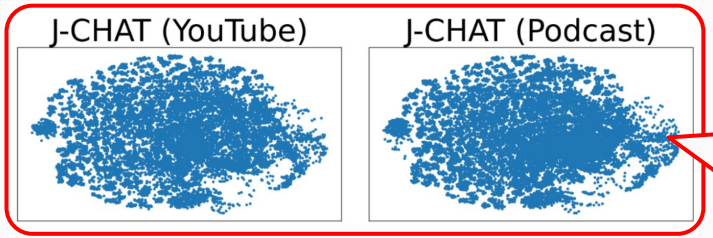
Table 2: Corpus statistics by its subsets, YouTube and Podcast. # means "number of".

feature	YouTube	Podcast	Total
total duration[hr]	11,001	57,891	68,892
# dialogue	1,013,488	3,924,009	4,937,497
mean duration [s]	39.07	53.11	50.23
mean # turns	7.58	10.68	10.10
mean # speakers	3.23	3.12	3.14

音韻的な分布 (HuBERT特徴量）

音声の音韻的特徴を表すHuBERTの分布を可視化

J-CHATはより多様な音韻的特徴を捉えたデータセット



J-CHATコーパスは
多様な音韻を網羅！

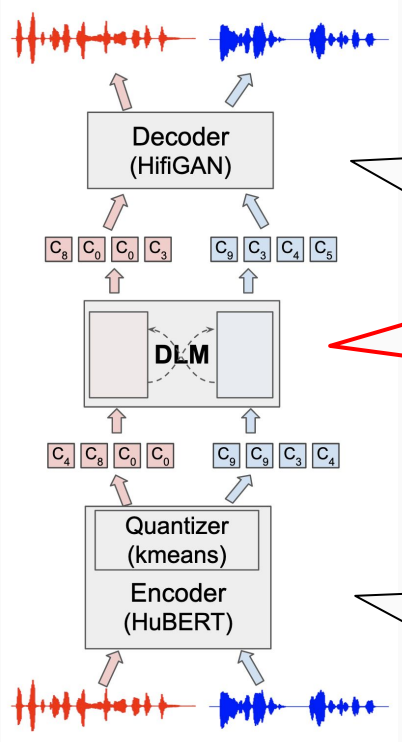
日本語
対話コーパス
[Saito+22]

日本語
非言語発話
[Xin+24]

対話音声合成による評価

モデル：dGSLM[Nguyen+22]

- Encoder・Decoderを用いて音声と離散トークン列を変換
  - これらは事前学習済みモデルを利用
- 実験では対話言語モデルを学習
- 学習には話者ごとに音声をわけておく必要がある
  - J-CHATには「いつ誰が話してるか」の情報が付随しているため，これを用いて話者交代時に音声をわける



対話音声合成による評価

- 実験条件：以下の条件で対話音声を合成
  - **resynth**: Encoder-Decoderを用いて原音声を再合成（比較用）
  - **dGSLM-YouTube**: J-CHATのYouTubeサブセットで学習
  - **dGSLM-podcast**: J-CHATのPodcastサブセットのみで学習
  - **dGSLM-J-CHAT**: J-CHATコーパス全体で学習
- 評価指標：以下の指標について，平均主観評価値 (MOS) を集計して評価
  - 有意味性MOS:「対話に意味があるか」
  - 自然性MOS:「人間らしいか」
    - 各指標につき, 60人の聴取者による評価実験を実施

結果1. ドメインの重要性

- 全体を用いることで，サブセットの場合よりも高い自然性・有意味性
- サブセット間では差が見られない（量が5倍程度異なるにも関わらず！）

→ 複数のドメインから収集することが対話音声合成に効果的！

※データが多様になるほど生成が困難になるはずであり，原因の調査は今後の課題

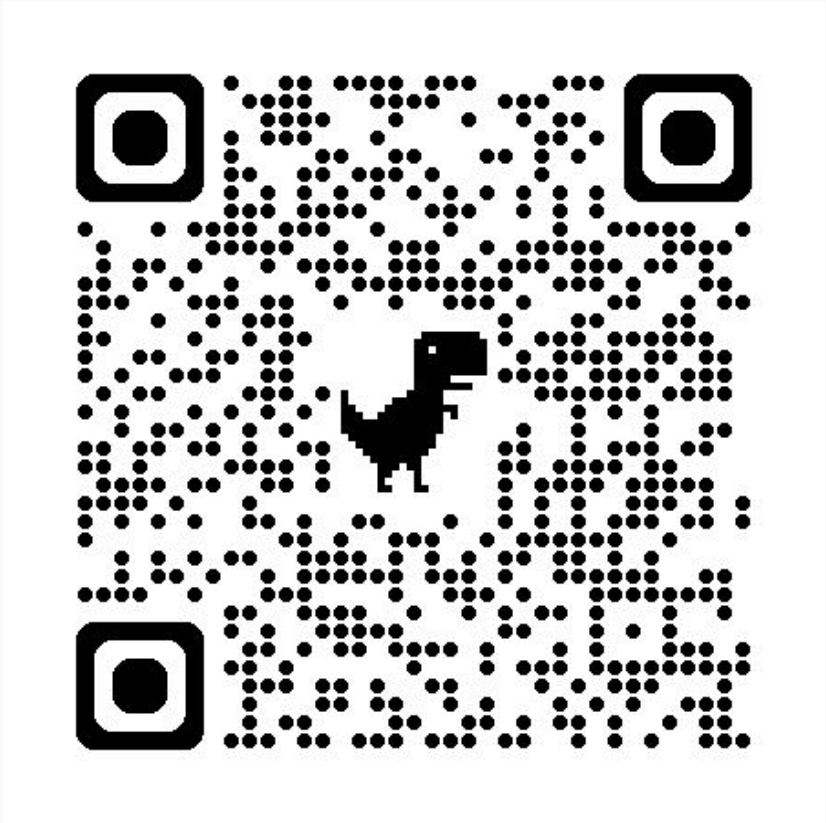
Model	Naturalness	Meaningfulness
resynth	2.55 ± 0.18	2.48 ± 0.18
dGSLM-Youtube	1.44 ± 0.13	1.56 ± 0.14
dGSLM-podcast	1.44 ± 0.13	1.52 ± 0.13
dGSLM-J-CHAT	2.28 ± 0.19	2.18 ± 0.19

結果2. 今後の課題：重複音声のモデル化

- 再合成の自然性・有意味性が低い（1↓～5↑の5段階評価で2.5程度）
- 原因：２人以上が同時に話す音声（重複音声）が混入したことでEncoder及びDecoderのモデル化が困難になったと考えられる
  - 従来は話者ごとにマイクをつけていたため問題にならなかった
  - J-CHATは全話者の音声を1つのマイクで拾っているため，分離できていない

Model	Naturalness	Meaningfulness
resynth	2.55 ± 0.18	2.48 ± 0.18
dGSLM-Youtube	1.44 ± 0.13	1.56 ± 0.14
dGSLM-podcast	1.44 ± 0.13	1.52 ± 0.13
dGSLM-J-CHAT	2.28 ± 0.19	2.18 ± 0.19

音声サンプル



まとめ

研究背景

- 音声対話言語モデルの学習に必要となる大規模な対話データセットの欠如

本研究の貢献

- 大規模日本語対話音声コーパス（J-CHAT）の構築
  - 日本語として最大・対話音声として最大

BigTech以外でも
研究の土俵に
上がれるように！

今後の予定

- J-CHAT を用いた対話モデリング手法の検討
  - 特に，従来とは異なり重複音声を異なるため，これに適した手法が必要

音声サンプル

Subset	resynth	dGSLM-YouTube	dGSLM-Podcast	dGSLM-J-CHAT
YouTube				
Podcast				

最初の5秒を元の音声でプロンプティング

resynth, dGSLM-J-CHAT は何かしら言語的な出力が得られている

背景：ロボットの社会実装事例の増加

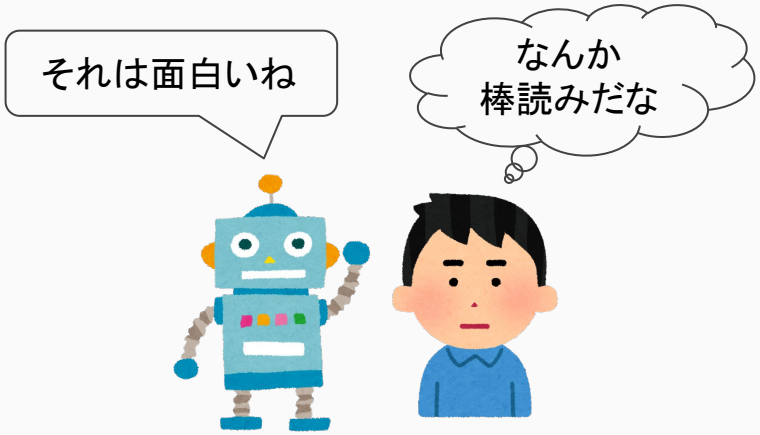
背景

- ロボット・AIが社会で利用されるように
  - 身近な例：ファミレスの配膳ロボット
- 他の多くのシーンでもロボットが増加
- 今後は「人間とロボットが共生する社会」に

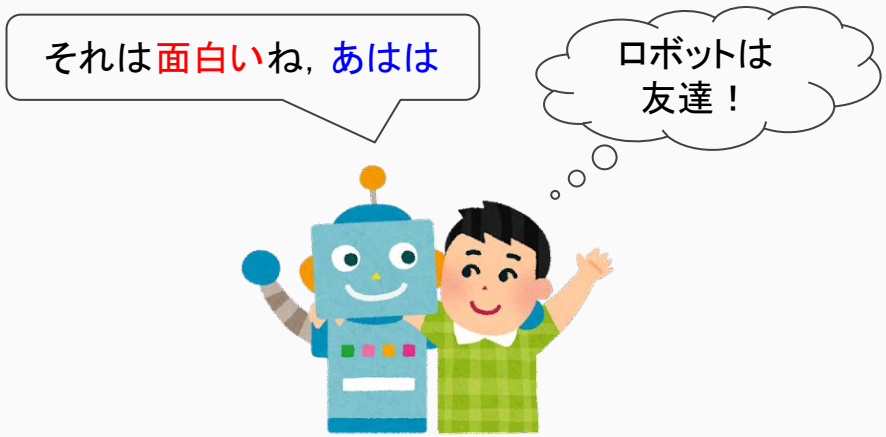
→ 人間とロボットのコミュニケーション手段が必要

配膳ロボット
（ファミレス等に導入）

目指すところ：表現力豊かな音声対話



従来のロボット



我々の目標

→ 表現力豊かな音声合成により人とロボットの共生を促進

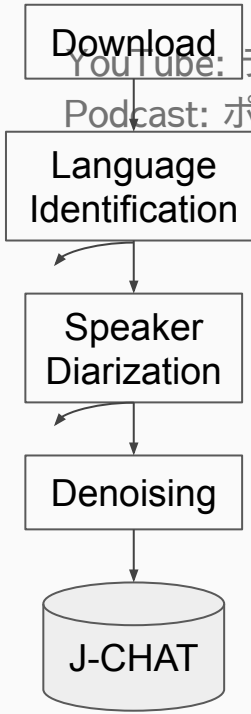
我々が解決したい社会課題

関連研究：対話音声言語モデル

dGSLM[Nguyen+22]

- 対話音声のみから言語モデルを学習
- ターンテイキングや笑い声などの生成を確認
- データセットが小さいため，有意味性が低い

データフィルタリングプロセス



YouTube: ランダムな単語でYouTube検索-> Download
Podcast: ポッドキャストチャンネルのRSSを収集 -> Download

Whisper[Radford+23]を用いて日本語を抽出
YouTube: 55.7% Podcast 84.7% を抽出

Pyannote diarizeを用いてDiarizationを実施
独話を除去．発話区間の80%以上を一人の話者が占めているデータを削除
YouTube 41.9% Podcast 45%を抽出

demucs[Rouard+23] を用いてBGMなどのノイズを除去

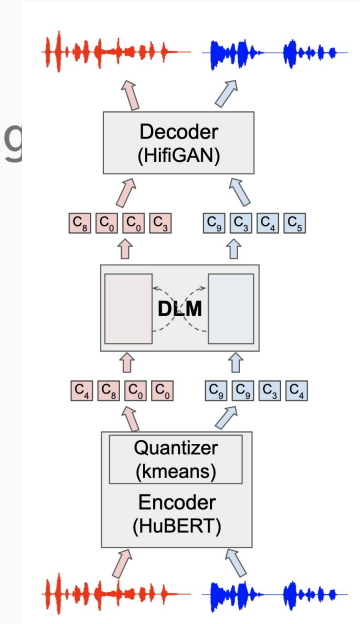
完成😻

対話音声合成による評価

モデル：dGSLM[Nguyen+22]

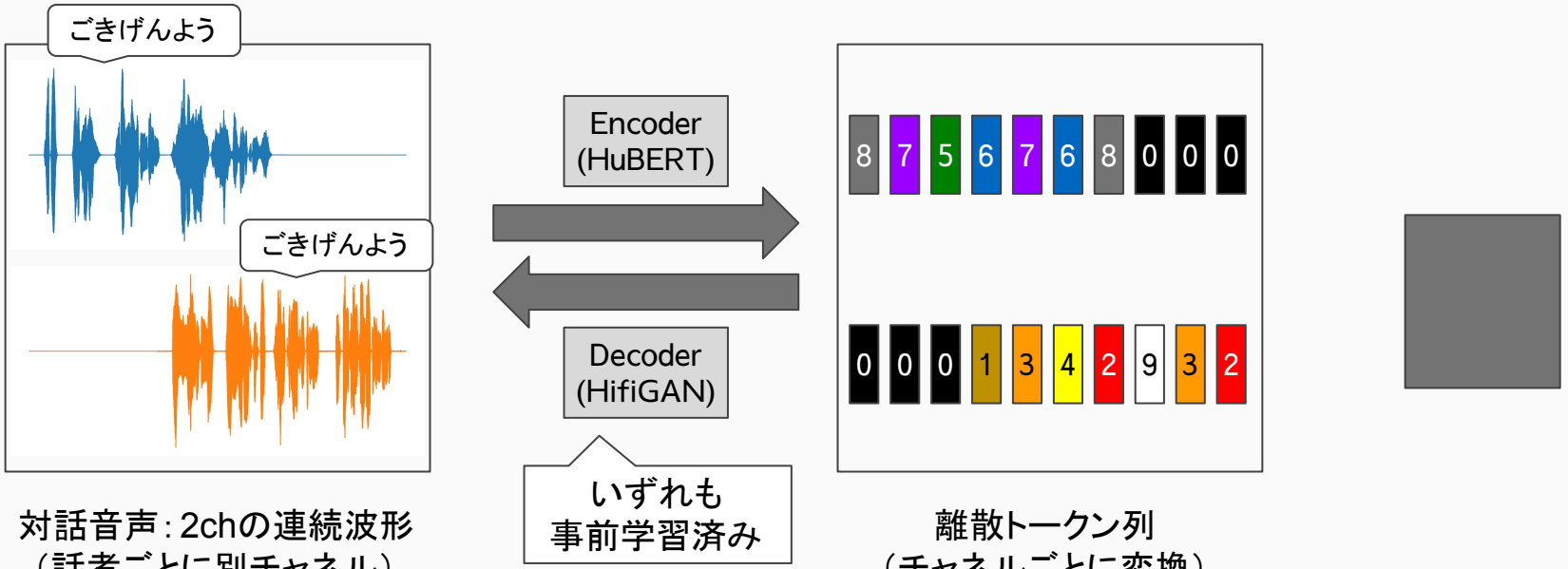
Diarization結果を元に作成

- チャネルごとに話者がわかれたデータでDialogue languag
- 実験条件
  - resynth DLMを用いず原音声から再合成
  - dGSLM-YouTube YouTubeのみで学習
  - dGSLM-podcast podcastのみで学習
  - dGSLM-J-CHAT 全部で学習
- 評価指標 一般的に使われている主観評価指標
  - 有意味性MOS「対話に意味があるか」
  - 自然性MOS 「人間らしいか」
  - 60人の聴取者による評価実験を実施



dGSLM[Nguyen+22] を用いた対話音声合成

対話音声合成手法 dGSLM[Nguyen+22]




0
1
2
3
4
5
6
7
8
9